



# Machine Learning in Action

Tatyana Goldberg

(goldberg@rostlab.org)

August 16, 2016

# Machine Learning in Biology

Beijing Genomics Institute in Shenzhen, China

June 2014



- **GenBank<sup>1</sup>**  
173,353,076 DNA sequences
- **UniProt<sup>2</sup>**  
69,560,473 protein sequences
- **GOLD<sup>3</sup>**  
total # of genomes 49,092

**Biological data is growing much faster than our knowledge of biological processes!**

1 <http://www.genomesonline.org/cgi-bin/GOLD/index.cgi>

2 <http://www.ncbi.nlm.nih.gov/genbank/>

3 <http://www.uniprot.org/>

# Naive Bayes

Car theft example: chances your car gets stolen?

Color	Brand	Age	Stolen?
Black	BMW	New	Yes
Black	BMW	Old	Yes
Green	BMW	New	Yes
Green	BMW	Old	No
Black	Lada	Old	No
Black	Lada	Old	No
Green	Lada	New	Yes
Green	Lada	New	No
Green	Lada	Old	No
Green	Lada	New	Yes



©<http://www.vistabmw.com>



©<http://ramario.nichost.ru>

Sample vectors  $x = (x_1, \dots, x_n) \in X$     Class labels  $y \in Y$

# Naive Bayes

Sample vectors  $x = (x_1, \dots, x_n) \in X$

Class labels  $y \in Y$

$$\underbrace{p(y|x)}_{\text{posterior}} = \frac{\overbrace{p(x|y)}^{\text{likelihood}} \cdot \overbrace{p(y)}^{\text{prior}}}{\underbrace{p(x)}_{\text{evidence}}}$$

$$p(x|y) = \prod_{i=1}^n p(x_i|y) \quad \text{„naïve“}$$

Thomas Bayes



©<http://www.wikipedia.org/>

# Naive Bayes

Sample vectors  $x = (x_1, \dots, x_n) \in X$

Class labels  $y \in Y$

$$\underbrace{p(y|x)}_{\text{posterior}} = \frac{\overbrace{p(x|y)}^{\text{likelihood}} \cdot \overbrace{p(y)}^{\text{prior}}}{\underbrace{p(x)}_{\text{evidence}}}$$

$$p(x|y) = \prod_{i=1}^n p(x_i|y) \quad \text{„naïve“}$$

Thomas Bayes



©<http://www.wikipedia.org/>

**Naive Bayes  
classifier adds a  
decision rule**



Decision rule

$$y_{max} = \arg \max_{y \in Y} p(x|y) \cdot p(y)$$

# Naive Bayes

Color	Brand	Age	Stolen?
Black	BMW	New	Yes
Black	BMW	Old	Yes
Green	BMW	New	Yes
Green	BMW	Old	No
Black	Lada	Old	No
Black	Lada	Old	No
Green	Lada	New	Yes
Green	Lada	New	No
Green	Lada	Old	No
Green	Lada	New	Yes
Black	Lada	New	?

$$p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)}$$

$$p(\text{Yes}) = 0.5$$

$$p(\text{No}) = 0.5$$

# Naive Bayes

Color	Brand	Age	Stolen?
Black	BMW	New	Yes
Black	BMW	Old	Yes
Green	BMW	New	Yes
Green	BMW	Old	No
Black	Lada	Old	No
Black	Lada	Old	No
Green	Lada	New	Yes
Green	Lada	New	No
Green	Lada	Old	No
Green	Lada	New	Yes
Black	Lada	New	?

$$p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)}$$

$$p(\text{Yes}) = 0.5 \quad p(\text{No}) = 0.5$$

$$p(\text{Black}|\text{Yes}) = \frac{2}{5}; \quad p(\text{Black}|\text{No}) = \frac{2}{5}$$

# Naive Bayes

Color	Brand	Age	Stolen?
Black	BMW	New	Yes
Black	BMW	Old	Yes
Green	BMW	New	Yes
Green	BMW	Old	No
Black	Lada	Old	No
Black	Lada	Old	No
Green	Lada	New	Yes
Green	Lada	New	No
Green	Lada	Old	No
Green	Lada	New	Yes
Black	Lada	New	?

$$p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)}$$

$$p(\text{Yes}) = 0.5 \quad p(\text{No}) = 0.5$$

$$p(\text{Black}|\text{Yes}) = \frac{2}{5}; \quad p(\text{Black}|\text{No}) = \frac{2}{5}$$

$$p(\text{Lada}|\text{Yes}) = \frac{2}{5}; \quad p(\text{Lada}|\text{No}) = \frac{4}{5}$$

# Naive Bayes

Color	Brand	Age	Stolen?
Black	BMW	New	Yes
Black	BMW	Old	Yes
Green	BMW	New	Yes
Green	BMW	Old	No
Black	Lada	Old	No
Black	Lada	Old	No
Green	Lada	New	Yes
Green	Lada	New	No
Green	Lada	Old	No
Green	Lada	New	Yes
Black	Lada	New	?

$$p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)}$$

$$p(\text{Yes}) = 0.5 \quad p(\text{No}) = 0.5$$

$$p(\text{Black}|\text{Yes}) = \frac{2}{5}; \quad p(\text{Black}|\text{No}) = \frac{2}{5}$$

$$p(\text{Lada}|\text{Yes}) = \frac{2}{5}; \quad p(\text{Lada}|\text{No}) = \frac{4}{5}$$

$$p(\text{New}|\text{Yes}) = \frac{4}{5}; \quad p(\text{New}|\text{No}) = \frac{1}{5}$$

# Naive Bayes

Color	Brand	Age	Stolen?
Black	BMW	New	Yes
Black	BMW	Old	Yes
Green	BMW	New	Yes
Green	BMW	Old	No
Black	Lada	Old	No
Black	Lada	Old	No
Green	Lada	New	Yes
Green	Lada	New	No
Green	Lada	Old	No
Green	Lada	New	Yes
Black	Lada	New	?

$$p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)}$$

$$p(\text{Yes}) = 0.5 \quad p(\text{No}) = 0.5$$

$$p(\text{Black}|\text{Yes}) = \frac{2}{5}; \quad p(\text{Black}|\text{No}) = \frac{2}{5}$$

$$p(\text{Lada}|\text{Yes}) = \frac{2}{5}; \quad p(\text{Lada}|\text{No}) = \frac{4}{5}$$

$$p(\text{New}|\text{Yes}) = \frac{4}{5}; \quad p(\text{New}|\text{No}) = \frac{1}{5}$$

$$p(\text{Yes}|\text{Black}, \text{Lada}, \text{New}) = \mathbf{0.064/z}$$

$$p(\text{No}|\text{Black}, \text{Lada}, \text{New}) = 0.032/z$$

# Naive Bayes

Color	Brand	Age	Stolen?
Black	BMW	New	Yes
Black	BMW	Old	Yes
Green	BMW	New	Yes
Green	BMW	Old	No
Black	Lada	Old	No
Black	Lada	Old	No
Green	Lada	New	Yes
Green	Lada	New	No
Green	Lada	Old	No
Green	Lada	New	Yes
Black	Lada	New	?
Green	BMW	New	?

$$p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)}$$

$$p(\text{Yes}) = 0.5 \quad p(\text{No}) = 0.5$$

$$p(\text{Black}|\text{Yes}) = \frac{2}{5}; \quad p(\text{Black}|\text{No}) = \frac{2}{5}$$

$$p(\text{Lada}|\text{Yes}) = \frac{2}{5}; \quad p(\text{Lada}|\text{No}) = \frac{4}{5}$$

$$p(\text{New}|\text{Yes}) = \frac{4}{5}; \quad p(\text{New}|\text{No}) = \frac{1}{5}$$

$$p(\text{Yes}|\text{Black}, \text{Lada}, \text{New}) = \mathbf{0.064/z}$$

$$p(\text{No}|\text{Black}, \text{Lada}, \text{New}) = 0.032/z$$

$$p(\text{Yes}|\text{Green}, \text{BMW}, \text{New}) = \mathbf{0.14/z}$$

$$p(\text{No}|\text{Green}, \text{BMW}, \text{New}) = 0.005/z$$

# Naive Bayes

Color	Brand	Age	Stolen?
Black	BMW	New	Yes
Black	BMW	Old	Yes
Green	BMW	New	Yes
Green	BMW	Old	No
Black	Lada	Old	No
Black	Lada	Old	No
Green	Lada	New	Yes
Green	Lada	New	No
Green	Lada	Old	No
Green	Lada	New	Yes
Black	Lada	New	?
Green	BMW	New	?
Green	Lada	?	?

$$p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)}$$

$$p(\text{Yes}) = 0.5 \quad p(\text{No}) = 0.5$$

$$p(\text{Black}|\text{Yes}) = \frac{2}{5}; \quad p(\text{Black}|\text{No}) = \frac{2}{5}$$

$$p(\text{Lada}|\text{Yes}) = \frac{2}{5}; \quad p(\text{Lada}|\text{No}) = \frac{4}{5}$$

$$p(\text{New}|\text{Yes}) = \frac{4}{5}; \quad p(\text{New}|\text{No}) = \frac{1}{5}$$

$$p(\text{Yes}|\text{Black}, \text{Lada}, \text{New}) = \mathbf{0.064/z}$$

$$p(\text{No}|\text{Black}, \text{Lada}, \text{New}) = 0.032/z$$

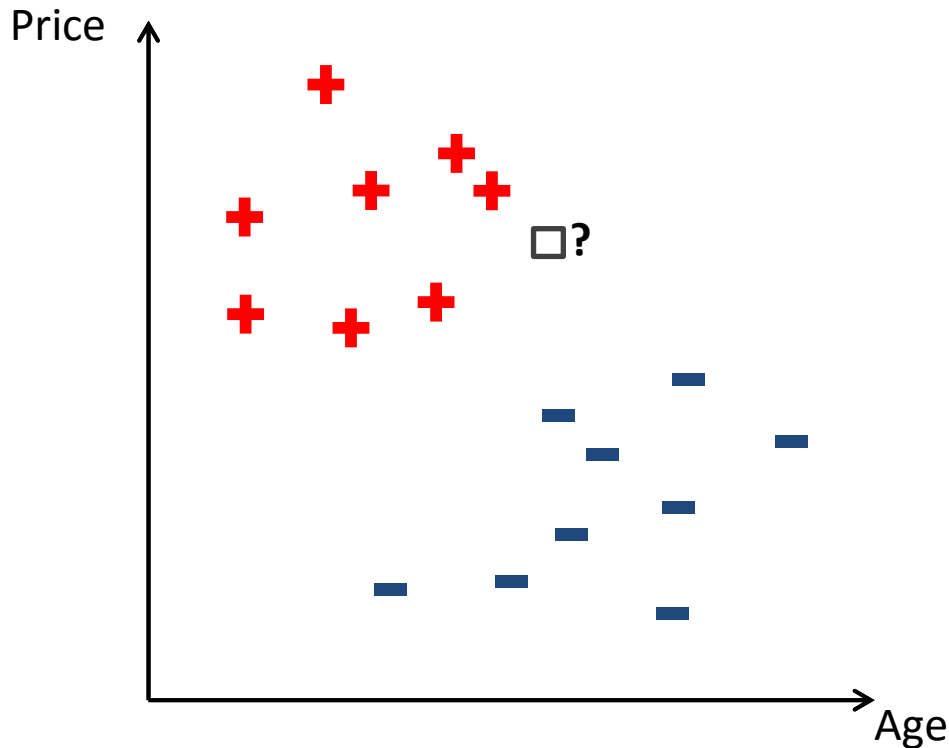
$$p(\text{Yes}|\text{Green}, \text{BMW}, \text{New}) = \mathbf{0.14/z}$$

$$p(\text{No}|\text{Green}, \text{BMW}, \text{New}) = 0.005/z$$

$$p(\text{Yes}|\text{Green}, \text{Lada}, ?) = 0.12/z$$

$$p(\text{No}|\text{Green}, \text{Lada}, ?) = \mathbf{0.24/z}$$

# Support Vector Machine



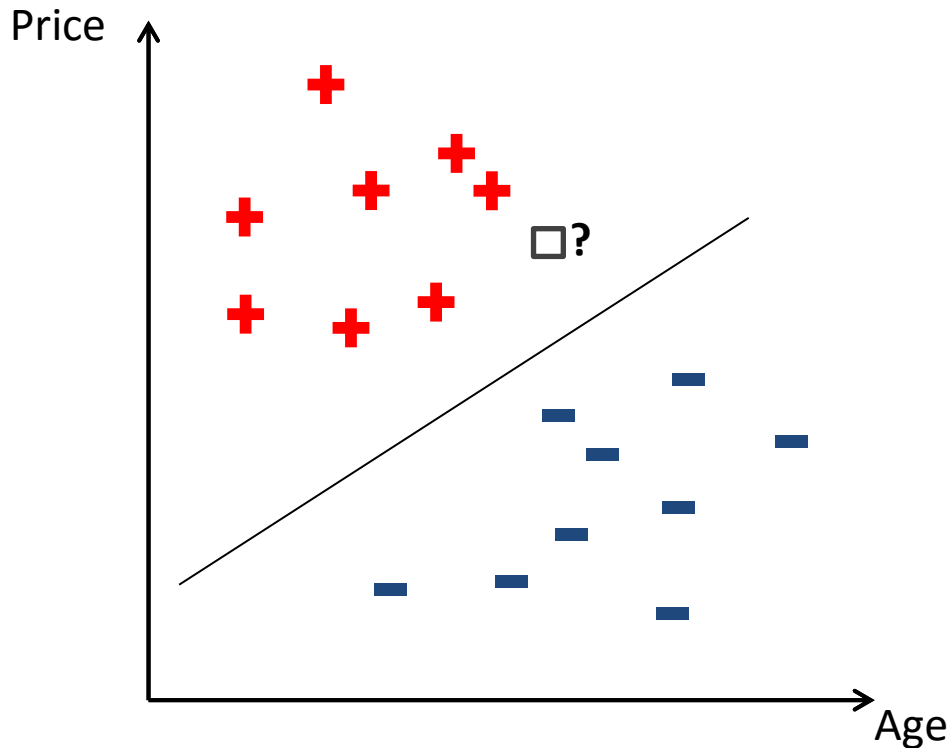
Vladimir Vapnik



©<http://www.nec-labs.com>

- Linear separation by building a hyperplane
- Hyperplane with the maximum margin is the best

# Support Vector Machine



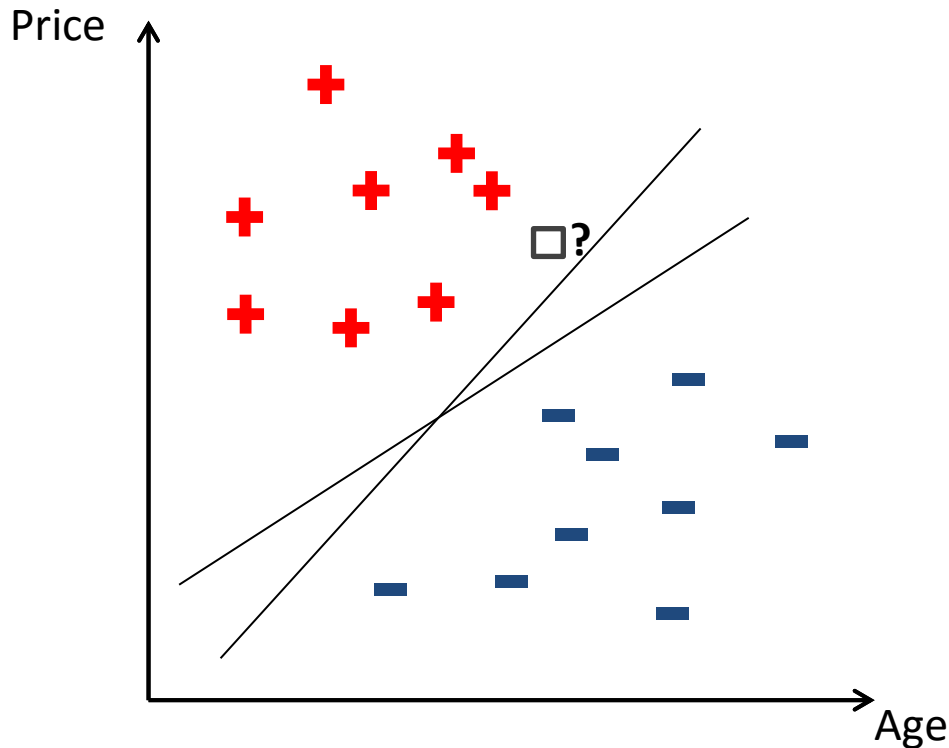
Vladimir Vapnik



©<http://www.nec-labs.com>

- Linear separation by building a hyperplane
- Hyperplane with the maximum margin is the best

# Support Vector Machine



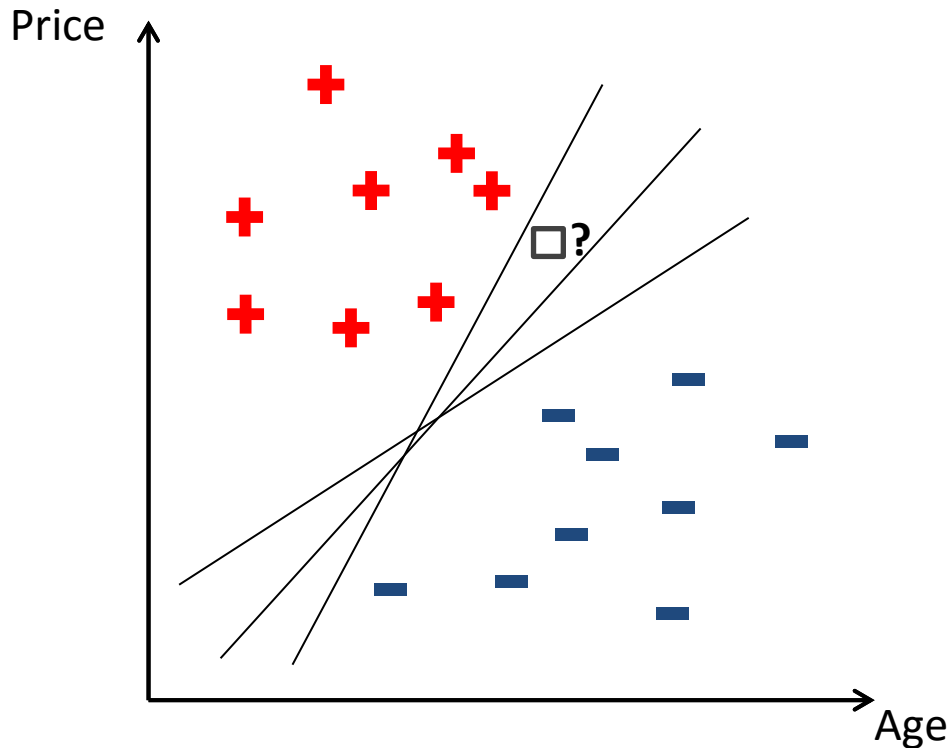
Vladimir Vapnik



©<http://www.nec-labs.com>

- Linear separation by building a hyperplane
- Hyperplane with the maximum margin is the best

# Support Vector Machine



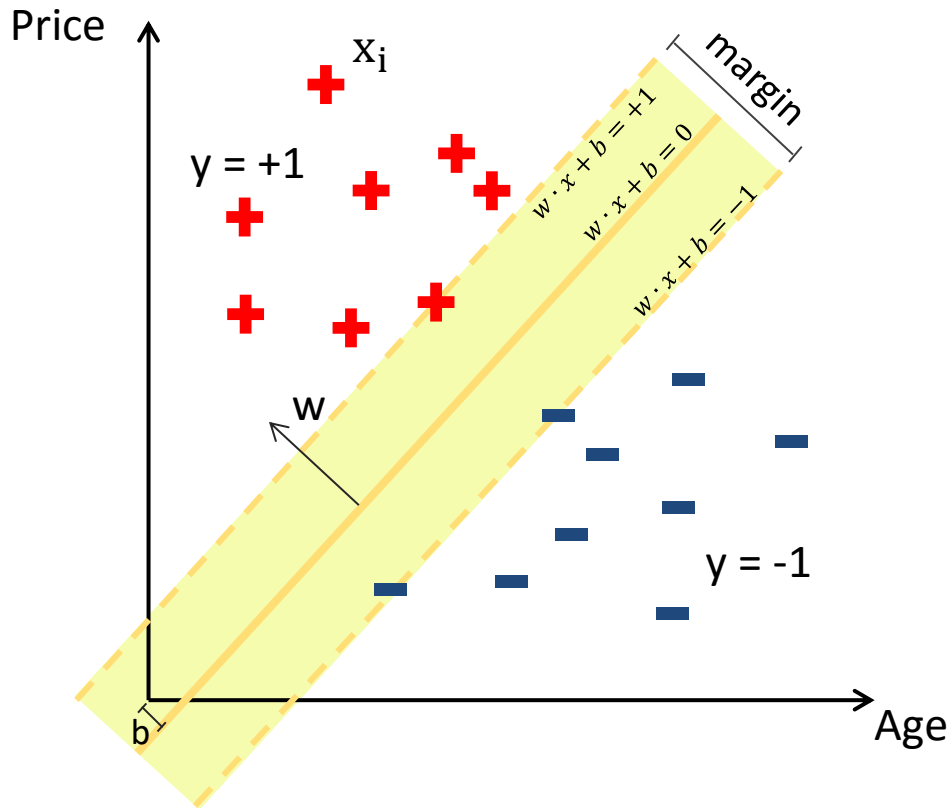
Vladimir Vapnik



©<http://www.nec-labs.com>

- Linear separation by building a hyperplane
- Hyperplane with the maximum margin is the best

# Support Vector Machine



Sample vectors  $\mathbf{x}_i = (x_1, \dots, x_n)_i,$   
 $i \in 1, \dots, N$

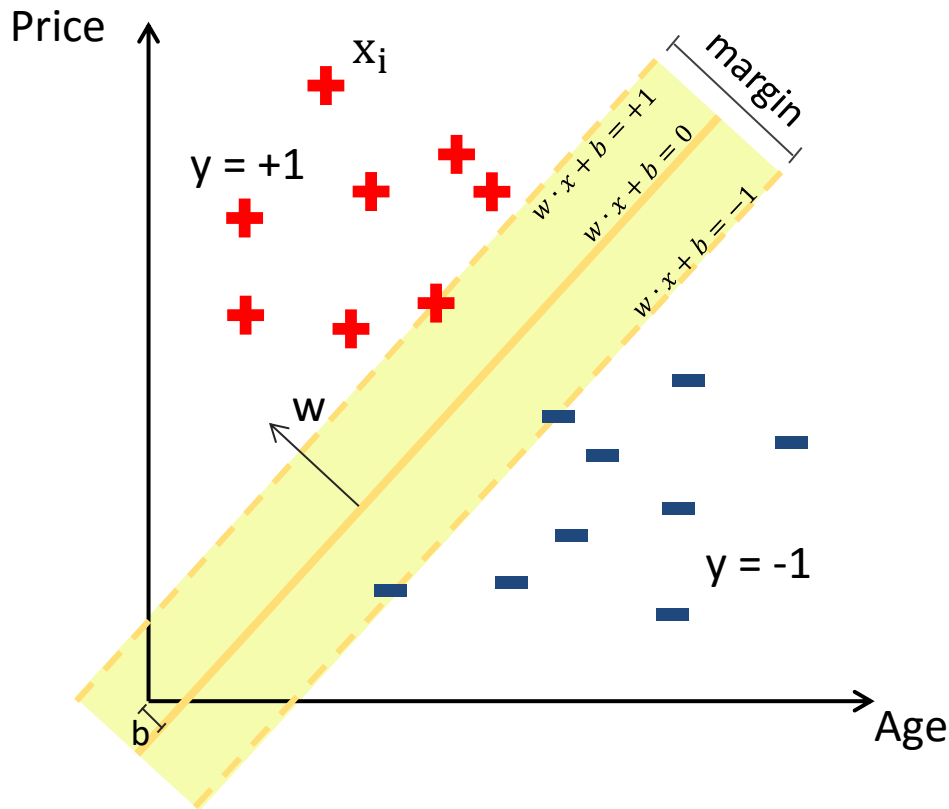
Class labels  $y_i \in \{-1, +1\}$

Separating hyperplane  $\mathbf{w}^T \mathbf{x}_i = 0$

$\mathbf{w}^T \mathbf{x}_i \geq 1$  for  $y_i = +1$

$\mathbf{w}^T \mathbf{x}_i \leq -1$  for  $y_i = -1$

# Support Vector Machine



Sample vectors  $\mathbf{x}_i = (x_1, \dots, x_n)_i,$   
 $i \in 1, \dots, N$

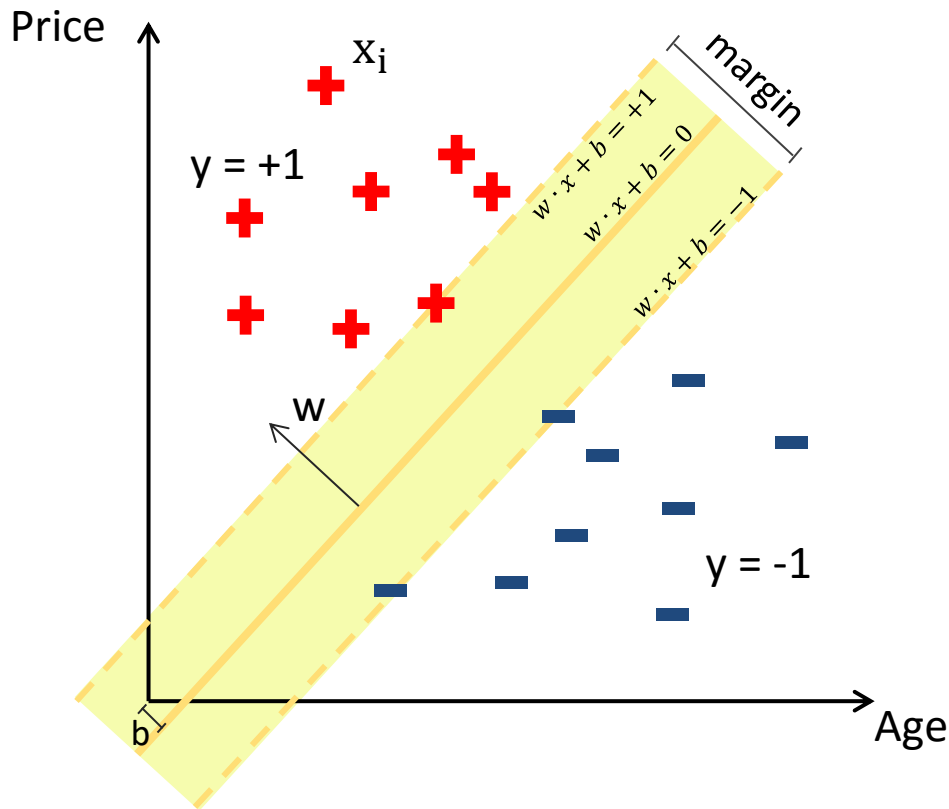
Class labels  $y_i \in \{-1, +1\}$

Separating hyperplane  $w^T \mathbf{x}_i + b = 0$

$w^T \mathbf{x}_i + b \geq 1$  for  $y_i = +1$

$w^T \mathbf{x}_i + b \leq -1$  for  $y_i = -1$

# Support Vector Machine



Sample vectors  $\mathbf{x}_i = (x_1, \dots, x_n)_i,$   
 $i \in 1, \dots, N$

Class labels  $y_i \in \{-1, +1\}$

Separating hyperplane  $w^T \mathbf{x}_i + b = 0$

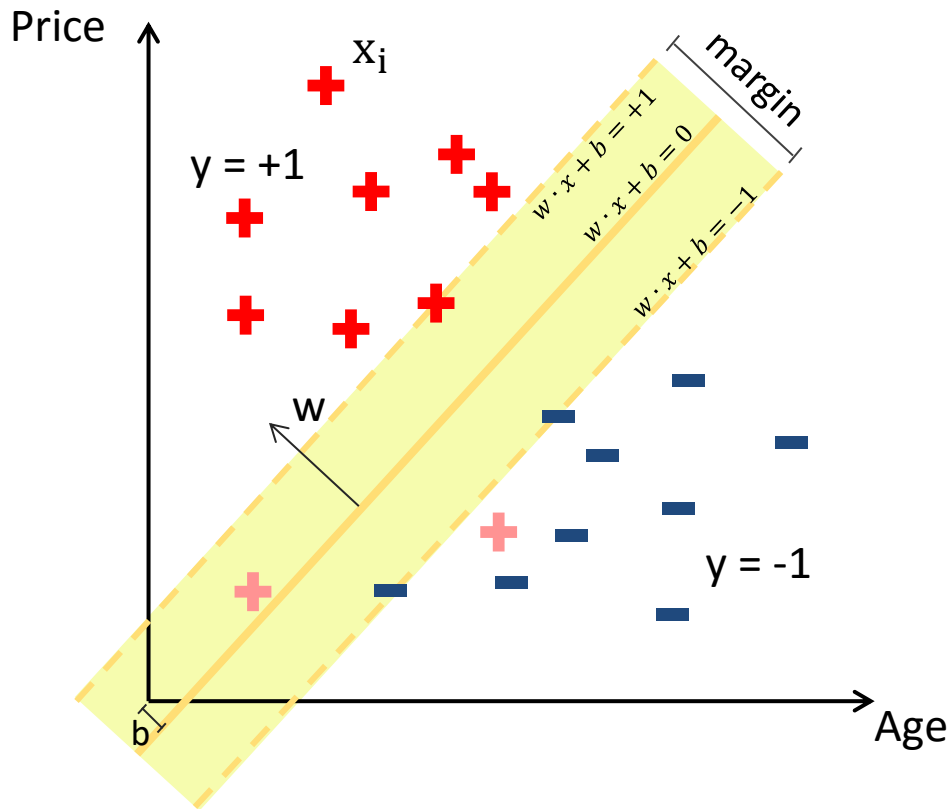
$w^T \mathbf{x}_i + b \geq 1$  for  $y_i = +1$

$w^T \mathbf{x}_i + b \leq -1$  for  $y_i = -1$

Can be combined to:

$y_i(w^T \mathbf{x}_i + b) \geq 1$  for  $\forall i$

# Support Vector Machine



Sample vectors  $\mathbf{x}_i = (x_1, \dots, x_n)_i,$   
 $i \in 1, \dots, N$

Class labels  $y_i \in \{-1, +1\}$

Separating hyperplane  $w^T \mathbf{x}_i + b = 0$

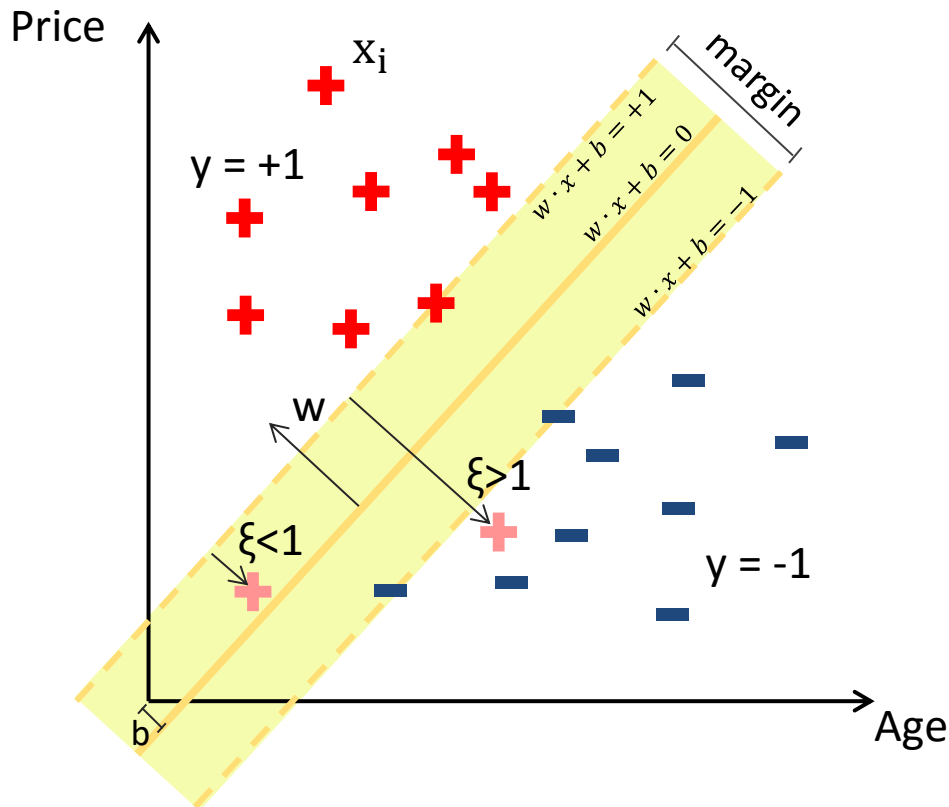
$w^T \mathbf{x}_i + b \geq 1$  for  $y_i = +1$

$w^T \mathbf{x}_i + b \leq -1$  for  $y_i = -1$

Can be combined to:

$y_i(w^T \mathbf{x}_i + b) \geq 1$  for  $\forall i$

# Support Vector Machine



Sample vectors  $\mathbf{x}_i = (x_1, \dots, x_n)_i,$   
 $i \in 1, \dots, N$

Class labels  $y_i \in \{-1, +1\}$

Separating hyperplane  $w^T \mathbf{x}_i + b = 0$

$w^T \mathbf{x}_i + b \geq 1$  for  $y_i = +1$

$w^T \mathbf{x}_i + b \leq -1$  for  $y_i = -1$

Can be combined to:

$y_i(w^T \mathbf{x}_i + b) \geq 1$  for  $\forall i$

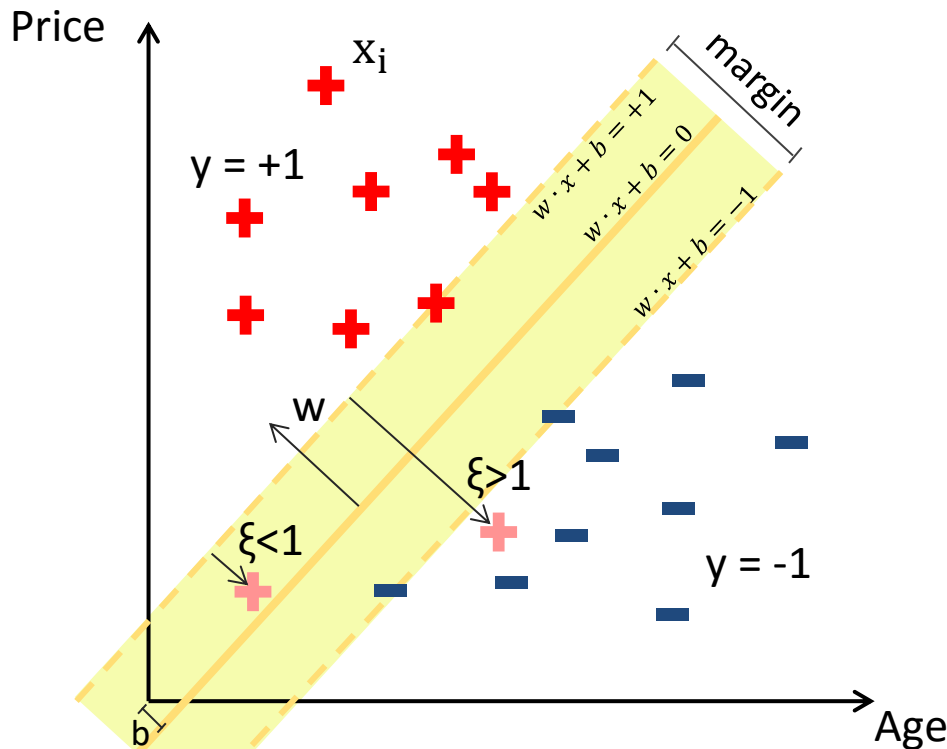
$y_i(w^T \mathbf{x}_i + b) \geq 1 - \xi_i$  for  $\forall i$

# Support Vector Machine



©[http://media.philly.com/images/And\\_then\\_a\\_miracle\\_happens\\_cartoon.jpg](http://media.philly.com/images/And_then_a_miracle_happens_cartoon.jpg)

# Support Vector Machine



Decision function

$$f(x) = \text{sgn}\left(\sum_{\forall i, \alpha_i > 0} \alpha_i y_i x_i^T x + b\right)$$

Support vectors  $(x_i | \alpha_i > 0)$

Sample vectors  $x_i = (x_1, \dots, x_n)_i,$   
 $i \in 1, \dots, N$

Class labels  $y_i \in \{-1, +1\}$

Separating hyperplane  $w^T x_i + b = 0$

$w^T x_i + b \geq 1$  for  $y_i = +1$

$w^T x_i + b \leq -1$  for  $y_i = -1$

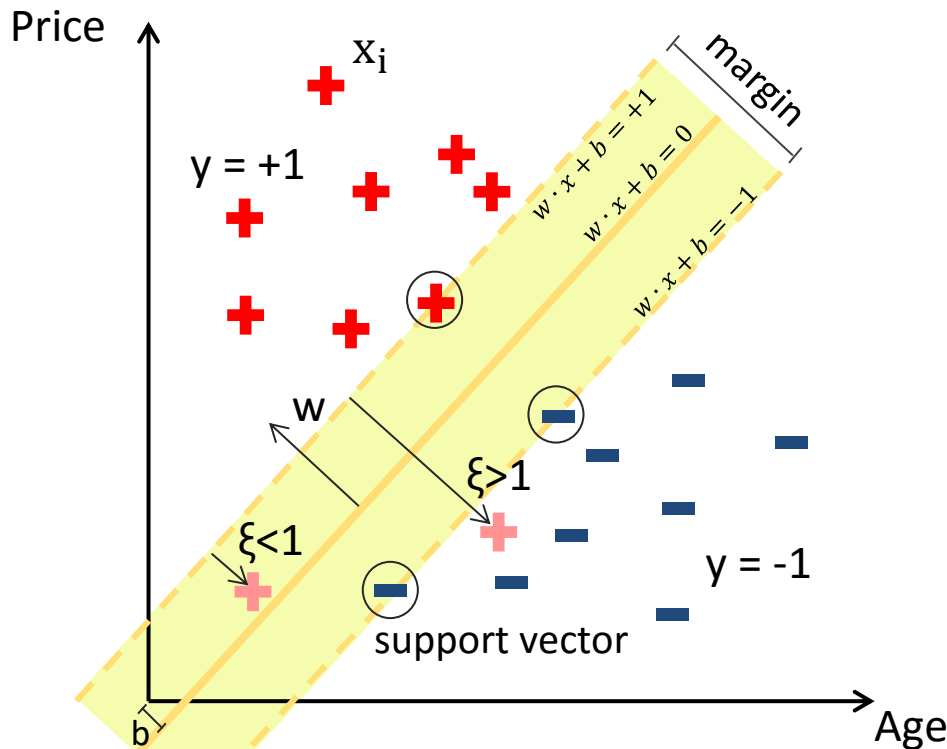
Can be combined to:

$y_i(w^T x_i + b) \geq 1$  for  $\forall i$

$y_i(w^T x_i + b) \geq 1 - \xi_i$  for  $\forall i$



# Support Vector Machine



Decision function

$$f(x) = \text{sgn}\left(\sum_{\forall i, \alpha_i > 0} \alpha_i y_i x_i^T x + b\right)$$

Support vectors  $(x_i | \alpha_i > 0)$

Sample vectors  $x_i = (x_1, \dots, x_n)_i,$   
 $i \in 1, \dots, N$

Class labels  $y_i \in \{-1, +1\}$

Separating hyperplane  $w^T x_i + b = 0$

$w^T x_i + b \geq 1$  for  $y_i = +1$

$w^T x_i + b \leq -1$  for  $y_i = -1$

Can be combined to:

$y_i(w^T x_i + b) \geq 1$  for  $\forall i$

$y_i(w^T x_i + b) \geq 1 - \xi_i$  for  $\forall i$



# Localization Predictions Indispensable

## High-throughput methods

- cost money
- not accurate
- not complete



Intracellular mitochondrial network, microtubules, nuclei

## Computational prediction

MTSHSYYKDRLGFDPEQQP  
GSNNMKRSSSRQTTHHHQ  
SYHHATTSSSQSPARISVSPG  
GNGTLEYQQVQRENNW...



Predictor



Protein Function

# The Signal Hypothesis

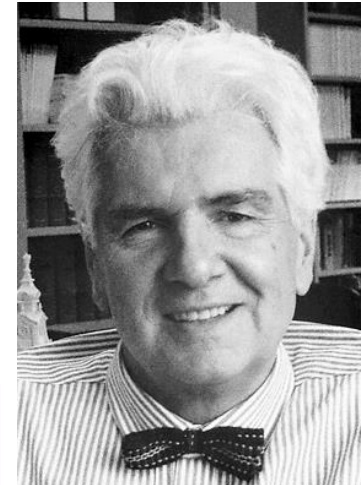


The Nobel Prize in Physiology or Medicine 1999

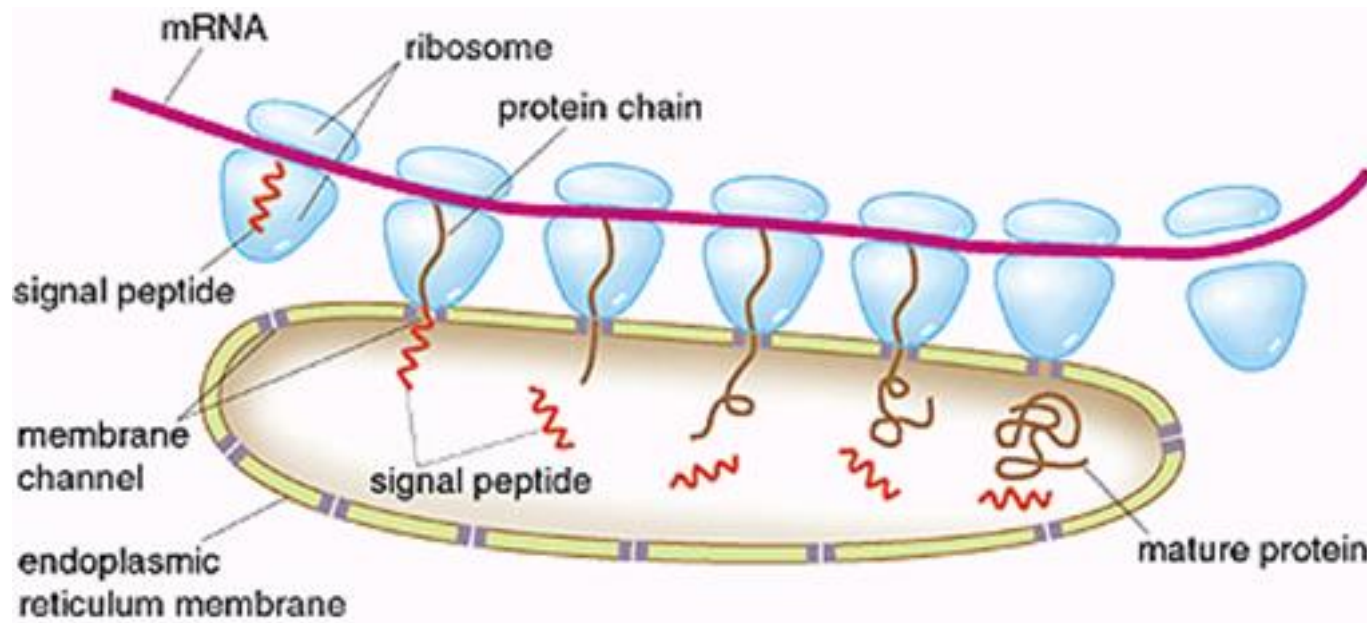
Günter Blobel

*„for the discovery that proteins have **intrinsic signals** that govern their transport and localization in the cell“*

Günter Blobel

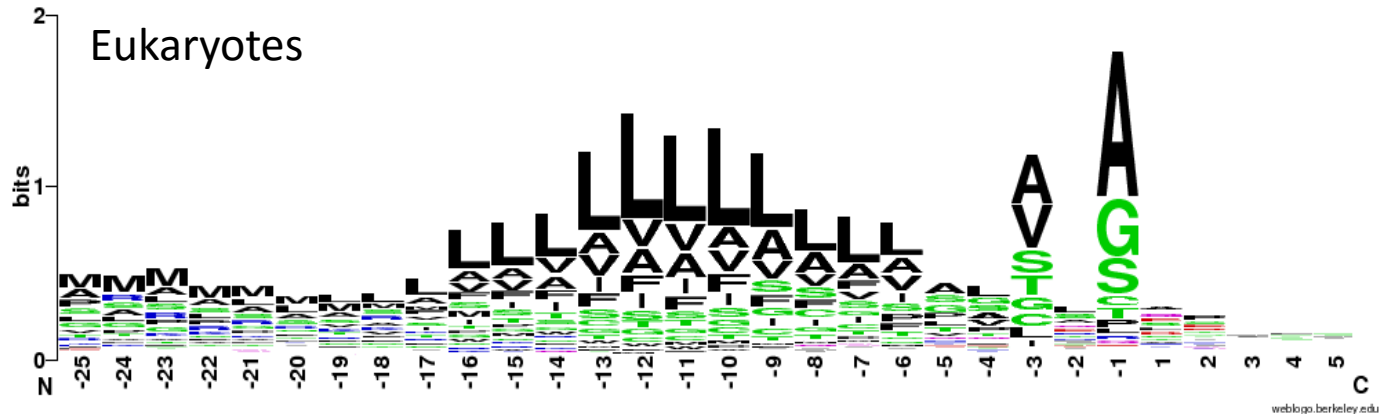


©<http://www.nobelprize.org/>



©<http://www.nobelprize.org/>

# Signal Peptide Prediction



Created with <http://weblogo.berkeley.edu/> using the SignapP 4.0 data set

- **SignalP v. 1.1 – 4.0:** prediction of signal peptides and their cleavage sites

# Henrik Nielsen



© <http://www.ebi.ac.uk>

# Gunnar von Heijne



© <http://www.sbc.su.se/>

Søren Brunak



© <http://www.cbs.dtu.dk>

# Data Preprocessing

- 40N terminal sequences of positive and negative samples

```

MKS LK S S T H D V P H P E H V V W A P P A Y D E Q H H L F F S H G T V L I G +1
M A Q S V T A F Q A A Y I S I E V L I A L V S V P G N I L V I W A V K M N Q A L +1
M G R K S L Y L L I V G I L I A Y Y I Y T P L P D N V E E P W R M M W I N A H L +1
M D N D G G A P P P P P T L V V E E P K K A E I R G V A F K E L F R F A D G L D +1
M G F E P L D W Y C K P V P N G V W T K T V D Y A F G A Y T P C A I D S F V L G +1
M A R S S L F T F L C L A V F I N G C L S Q I E Q Q S P W E F Q G S E V W Q Q H -1
M A V M A P R T L V L L L S G A L A L T Q T W A G S H S M R Y F Y T S V S R P G -1
M V E M L P T V A V L V L A V S V V A K D N T T C D G P C G L R F R Q N S Q A G -1
M N S N L P A E N L T I A V N M T K T L P T A V T H G F N S T N D P P S M S I T -1
M A L H M I L V M V S L L P L L E A Q N P E H V N I T I G D P I T N E T L S W L -1
M A N K L F L V S A T L A F F F L L T N A S I Y R T I V E V D E D D A T N P A G -1

```

- Convert sequences into **feature vectors**  
(e.g. amino acid composition)

Sequence

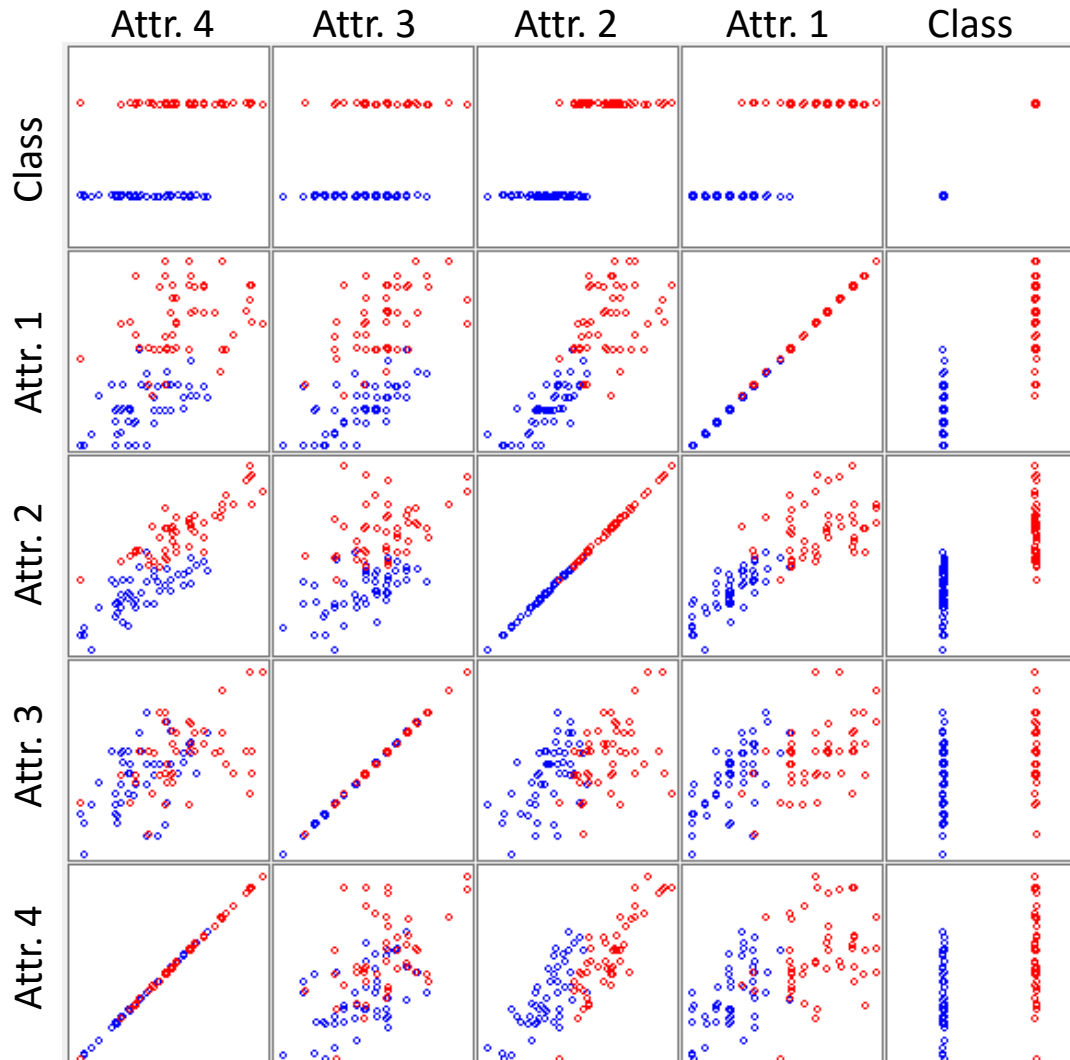
„MPPPAD“



20-dimensional feature vector

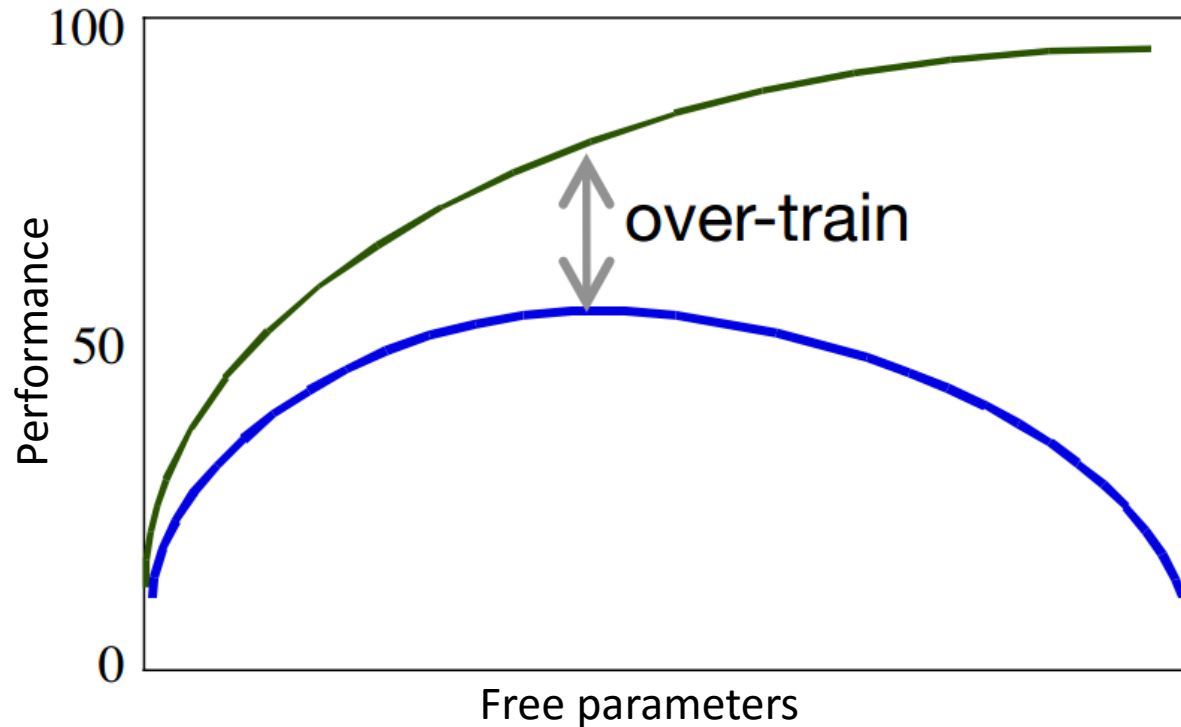
[1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 3, 0, 0, 0, 0, 0]

# Visualize Your Data



- 2D Scatterplots for all attribute pairs
- Colors are the labels (+1/-1)
- Some clusters are easily separable (class/any attribute, Attr.1 /Attr. 3)
- Whereas clusters of Attr3/Attr4 are much harder to separate

# Overfitting: theory

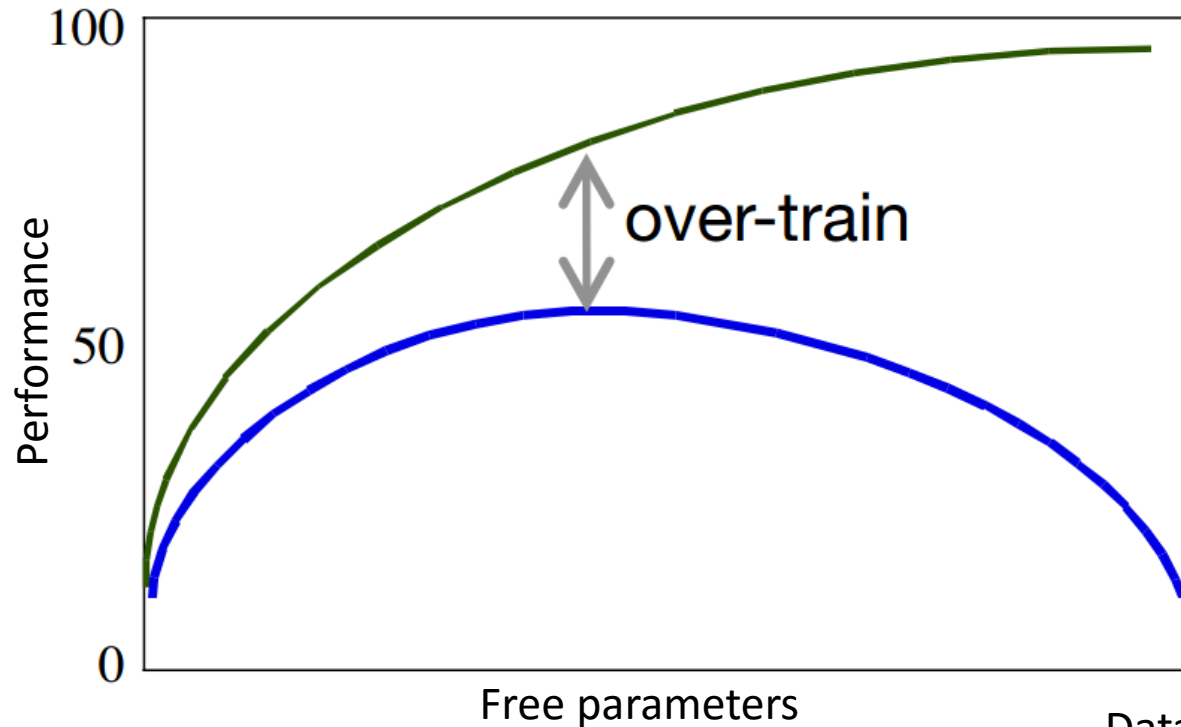


© PredictProtein lecture ,  
Prof. Rost

- The predictor fits the data too well
- Many more free parameters than samples
- Poor prediction on new data sets
- Rule of thumb:  $\text{samples} > 10 \cdot \text{free parameters}$

➡ Check for overfitting using cross-validation

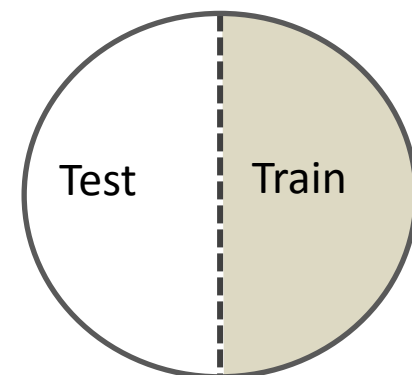
# Overfitting: theory



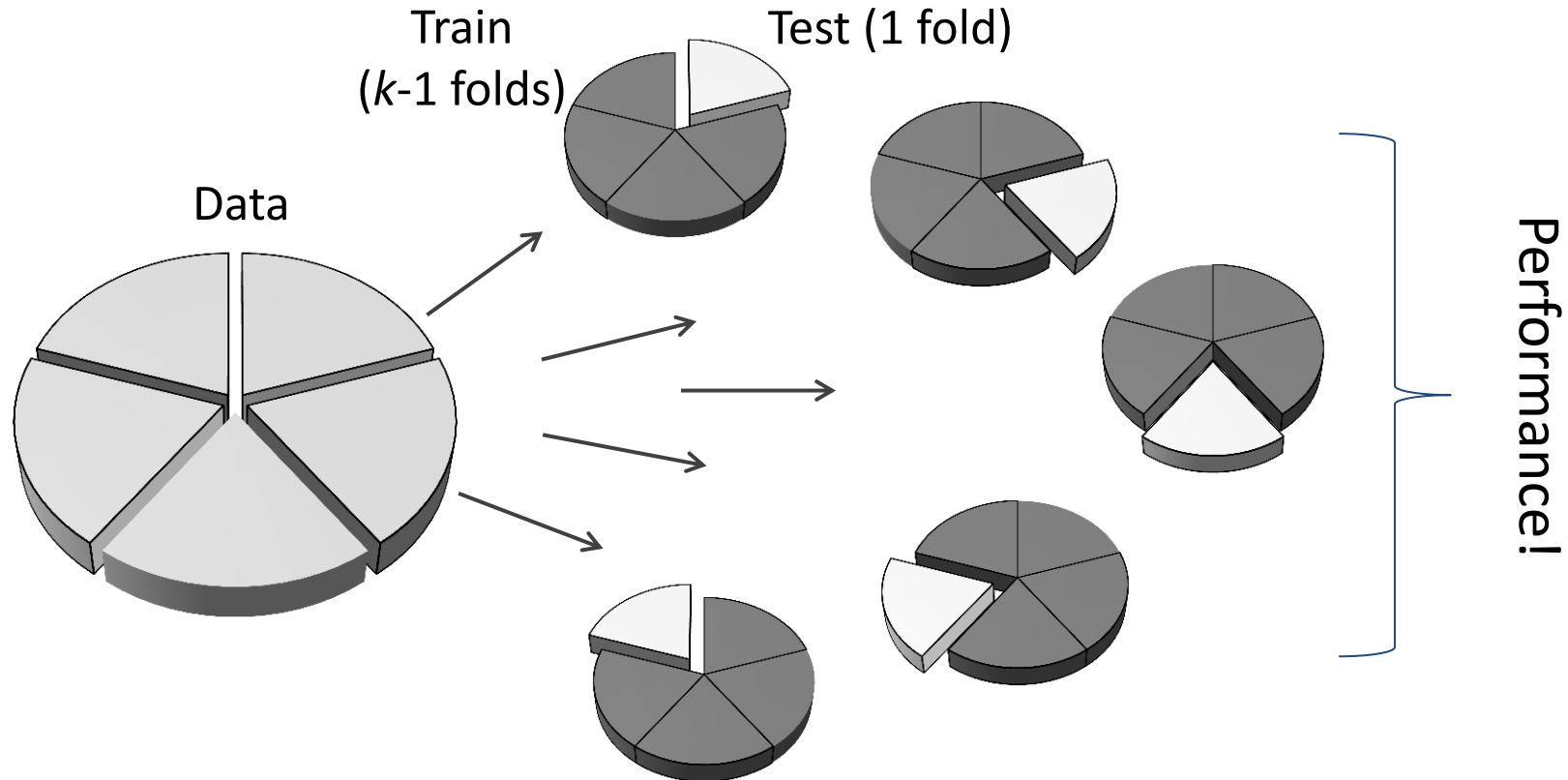
© PredictProtein lecture ,  
Prof. Rost

- The predictor fits the data too well
- Many more free parameters than samples
- Poor prediction on new data sets
- Rule of thumb:  $\text{samples} > 10 \cdot \text{free parameters}$

➡ Check for overfitting using cross-validation



# Stratified k-fold Cross-validation



*K-folds*: a random partitioning into  $k$  equally sized subsets, use each subset for testing exactly once and the remaining  $k-1$  subsets for training

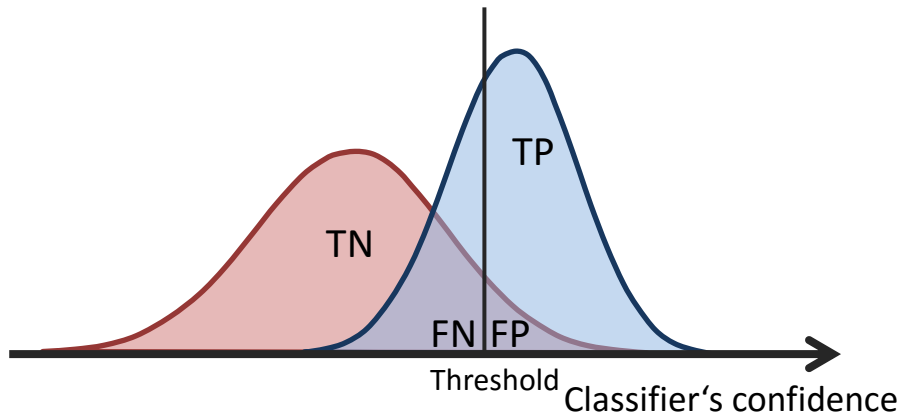
*Stratified*: each fold has the same proportion of each class value

# Performance Metric: AUC

		predicted	
		+1	-1
observed	+1	TP	FN
	-1	FP	TN

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

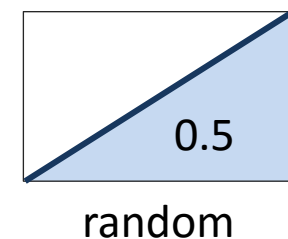
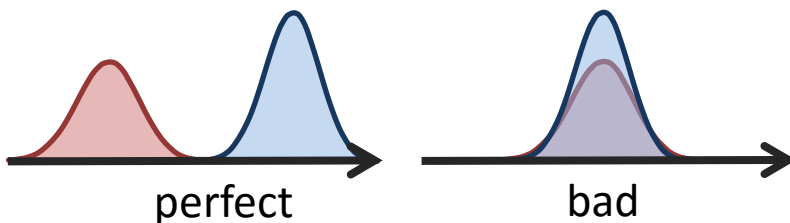
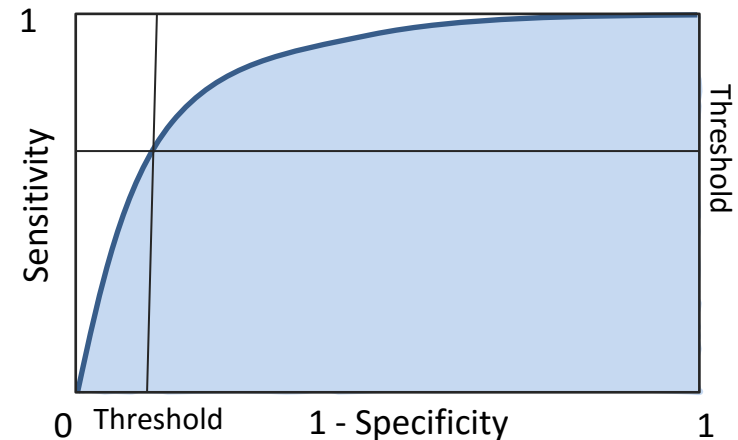
$$\text{Specificity} = \frac{TN}{TN+FP}$$



**ROC:** Receiver Operator Characteristic

**AUC:** Area under the ROC Curve

- Threshold independent
- Better than Acc, Cov, F1, etc.



# WEKA

## What is Weka?

Weka is a bird found only in New Zealand



Ian H. Witten



© <http://www.sai.com.ar>

## *Waikato Environment for Knowledge Analysis*

- Machine learning workbench for data mining tasks
  - 100+ algorithms for classification
  - 75+ for data pre-processing and analysis
  - 20+ for clustering, finding association rules, etc.
  - 25 to assist with feature selection
- <http://www.cs.waikato.ac.nz/ml/weka/>
- Java-based & supports multiple platforms

Eibe Frank



© <http://arnetminer.org>



THE UNIVERSITY OF  
**WAIKATO**  
*Te Whare Wānanga o Waikato*

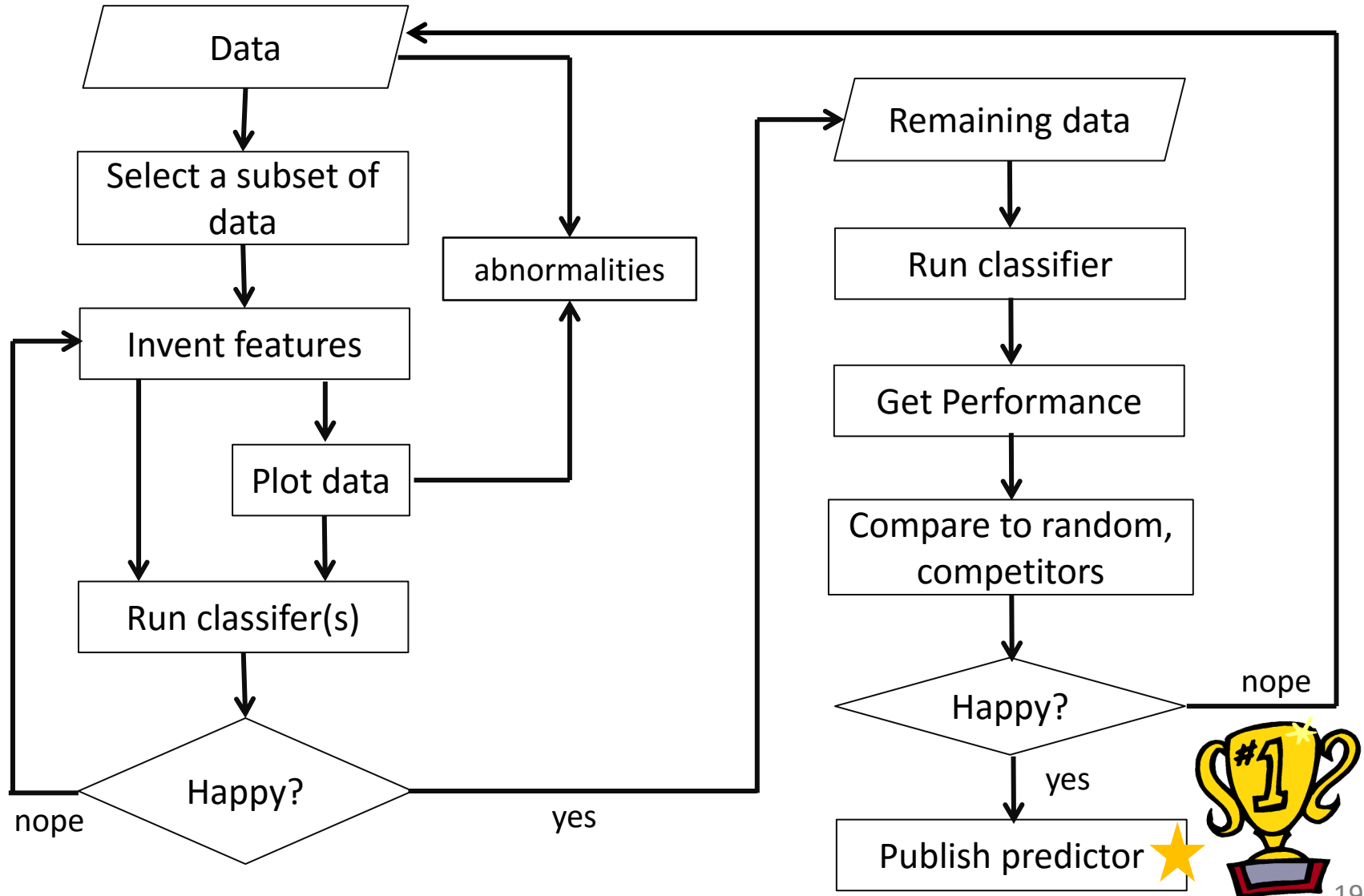
# WEKA in Action

---



Live Demo  
Presentation

# ML Workflow



# LOCTREE 3

## Protein Subcellular Localization Prediction System

Protein ID	Score	Expected Accuracy	Localization Class	Gene Ontology Terms	Annotation Type
E9PX37_MOUSE	99	98%	secreted	extracellular region GO:0005576;	LOCTREE2

Predicted Localization: secreted

>E9PX37\_MOUSE

